日　本　国　特　許　庁
JAPAN　PATENT　OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application:　２００２年　８月２９日

出願番号
Application Number:　特願２００２－２５０２８１

[ST.10/C]:　[ＪＰ２００２－２５０２８１]

出願人
Applicant(s):　株式会社リコー

２００３年　７月　３日

特許庁長官
Commissioner,
Japan Patent Office　太田信一郎

出証番号　出証特２００３－３０５２８４８

【書類名】　特許願
【整理番号】　0203184
【提出日】　平成14年 8月29日
【あて先】　特許庁長官　　太田　信一郎　殿
【国際特許分類】　G06F 17/30
【発明の名称】　単語出現度計算装置、文書検索装置、キーワード抽出装置、文書要約装置、文書分類装置、プログラム及び記憶媒体
【請求項の数】　20
【発明者】
　【住所又は居所】　東京都大田区中馬込１丁目３番６号　株式会社リコー内
　【氏名】　真野　博子
【発明者】
　【住所又は居所】　東京都大田区中馬込１丁目３番６号　株式会社リコー内
　【氏名】　伊東　秀夫
【特許出願人】
　【識別番号】　000006747
　【氏名又は名称】　株式会社リコー
　【代表者】　桜井　正光
【代理人】
　【識別番号】　100101177
　【弁理士】
　【氏名又は名称】　柏木　慎史
　【電話番号】　03(5333)4133
【選任した代理人】
　【識別番号】　100102130
　【弁理士】
　【氏名又は名称】　小山　尚人
　【電話番号】　03(5333)4133

出証特２００３－３０５２８４８

【選任した代理人】

【識別番号】　100072110

【弁理士】

【氏名又は名称】　柏木　明

【電話番号】　03(5333)4133

【手数料の表示】

【予納台帳番号】　063027

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9808802

【包括委任状番号】　0004335

【プルーフの要否】　要

【書類名】　　明細書

【発明の名称】　単語出現度計算装置、文書検索装置、キーワード抽出装置、文書要約装置、文書分類装置、プログラム及び記憶媒体

【特許請求の範囲】

【請求項１】　文字列の入力を受付ける入力手段と、

この受付けた文字列から単語を抽出する単語抽出手段と、

この抽出した各単語について所定の文書群の各文書における特定の部位での出現の度合いを計算する出現度計算手段と、

を備えている単語出現度計算装置。

【請求項２】　前記出現度計算手段は、前記特定の部位を前記文書における見出しとする、請求項１に記載の単語出現度計算装置。

【請求項３】　前記出現度計算手段は、前記特定の部位を前記文書における要約とする、請求項１に記載の単語出現度計算装置。

【請求項４】　前記出現度計算手段は、前記特定の部位を前記文書における見出し及び要約とする、請求項１に記載の単語出現度計算装置。

【請求項５】　前記出現度計算手段は、“前記文書群で前記単語が前記見出しで出現する文書数 ／ 前記文書群で前記単語の出現する全文書数”を計算することにより前記出現の度合いを計算する、請求項２に記載の単語出現度計算装置。

【請求項６】　前記出現度計算手段は、“前記文書群で前記単語が前記要約で出現する文書数 ／ 前記文書群で前記単語の出現する全文書数”を計算することにより前記出現の度合いを計算する、請求項３に記載の単語出現度計算装置。

【請求項７】　前記出現度計算手段は、“前記文書群で前記単語が前記見出し又は前記要約で出現する文書数 ／ 前記文書群で前記単語の出現する全文書数”を計算することにより前記出現の度合いを計算する、請求項４に記載の単語出現度計算装置。

【請求項８】　前記出現度計算手段は、“（単語が見出しで出現する文書数 ／ 単語の出現する全文書数） ＋ （単語が要約で出現する文書数 ／ 単語の出現する全文書数）”を計算することにより前記出現の度合いを計算する、請求項

４に記載の単語出現度計算装置。

【請求項９】　前記特定の部位の種類について選択を受付ける選択手段を備え、

前記出現度計算手段は、この選択された特定の部位における前記出現の度合いを計算する、請求項１～８の何れかの一に記載の単語出現度計算装置。

【請求項１０】　請求項１～９の何れかの一に記載の単語出現度計算装置と、

前記出現の度合いを利用して前記文書から検索語を選出する検索語選出手段と、

この選出した検索語に適合する文書を前記文書群から選出する文書選出手段と、

を備えている文書検索装置。

【請求項１１】　請求項１～９の何れかの一に記載の単語出現度計算装置と、

前記出現の度合いを利用して前記文書からキーワードを抽出するキーワード抽出手段と、

を備えているキーワード抽出装置。

【請求項１２】　請求項１～９の何れかの一に記載の単語出現度計算装置と、

前記出現の度合いを利用して前記文書からキーワードを抽出するキーワード抽出手段と、

前記出現の度合いを利用して前記文書から文を抽出して要約文とする要約作成手段と、

を備えている文書要約装置。

【請求項１３】　請求項１～９の何れかの一に記載の単語出現度計算装置と、

前記出現の度合いを利用して前記文書から分類キーワードを抽出する分類キーワード抽出手段と、

この抽出結果を利用して前記文書群の文書を分類する分類手段と、

を備えている文書分類装置。

【請求項１４】　文字列の入力を受付ける入力手段と、

この受付けた文字列から単語を抽出する単語抽出手段と、

この抽出した各単語について所定の第１の文書群の各文書における出現の度合いを計算する出現度計算手段と、

この出現の度合いを利用して前記文書から検索語を選出する検索語選出手段と、

この選出した検索語に適合する文書を前記第１の文書群とは別の第２の文書群から選出する文書選出手段と、

を備えている文書検索装置。

【請求項１５】　請求項１～９の何れかの一に記載の単語出現度計算装置の各手段が実行する処理をコンピュータに実行させるコンピュータに読取り可能なプログラム。

【請求項１６】　請求項１０又は１５に記載の文書検索装置の各手段が実行する処理をコンピュータに実行させるコンピュータに読取り可能なプログラム。

【請求項１７】　請求項１１に記載のキーワード抽出装置の各手段が実行する処理をコンピュータに実行させるコンピュータに読取り可能なプログラム。

【請求項１８】　請求項１２に記載の文書要約装置の各手段が実行する処理をコンピュータに実行させるコンピュータに読取り可能なプログラム。

【請求項１９】　請求項１３に記載の文書分類装置の各手段が実行する処理をコンピュータに実行させるコンピュータに読取り可能なプログラム。

【請求項２０】　請求項１５～１９の何れかの一に記載のプログラムを記憶している記憶媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、文書から単語を抽出してその単語の所定の文書群における出現度合いを計算する、単語出現度計算装置、文書検索装置、キーワード抽出装置、文書要約装置、文書分類装置、プログラム及び記憶媒体に関する。

【０００２】

【従来の技術】

文書を多数集積している文書データベースからユーザーの必要とする文書を探し出すには、ユーザーがひとつあるいは数個程度の単語からなるキーワードを入力し、そのキーワードに適合する文書を選出する方法が一般的である。しかし、ユーザーの利用目的によっては、単語でなく、文を検索要求としたい場合もある。検索要求が短文２～３文程度であれば、検索要求から助詞などの不要語を取り除いて検索語とすれば、ユーザーのもとめる文書を充分な検索精度で探し出すことができる。たとえば、特開2001-142897号公報では、検索要求から複数単語の連続を抽出し検索する方法が提案されている。

【０００３】

【発明が解決しようとする課題】

しかしながら、もっと長い検索要求、例えば、文書全体があたえられたような場合には、この方法では、検索語が多くなりすぎ、検索に多大な時間がかかるだけでなく、ノイズの多い検索となり、検索精度が低下することが多い。

【０００４】

例えば、「昨年」「一昨年」などの副詞的名詞は、ほとんどの場合、検索に有用でないが、こういった単語は、取り除かれる不要語としてもれなく定義するのが難しいという不具合がある。

【０００５】

また、長い検索要求でも許容するようになると、短いキーワードによる入力では比較的問題にならなかった、文体や語彙や内容領域が検索におよぼす影響が大きくなり、特に、検索対象文書と大きく異なる文体や語彙や内容領域を持つ検索要求が入力された場合、例えば、新聞記事を検索要求として特許公報を検索対象文書とするような場合には、検索精度の低下が見られるという不具合がある。例を挙げると、「発売」などの単語は、新聞記事には多くみられても特許公報に出てくることは少ないが、検索では一般に検索対象文書の文書データベースでの出現文書数の少ない単語を重要とみなすので、「発売」は重要語とみなされることになってしまう。

【０００６】

本発明の目的は、長い文書が入力された場合でも、文書検索等に有用な重要語のみを選出できるようにすることである。

【０００７】

また、別の目的は、検索対象等となる文書群と大きく異なる文体や語彙や内容領域を持つ文章が入力された場合でも、適切な単語が選出されるようにすることである。

【０００８】

【課題を解決するための手段】

請求項１に記載の発明は、文字列の入力を受付ける入力手段と、この受付けた文字列から単語を抽出する単語抽出手段と、この抽出した各単語について所定の文書群の各文書における特定の部位での出現の度合いを計算する出現度計算手段と、を備えている単語出現度計算装置である。

【０００９】

したがって、文字列から抽出した単語の有用度を決定するのに、それぞれの単語が所定の文書群の中で出現する部位に着目し、その単語が出現する全出現数に比して所定の重要な部位に出現する度合いを計算することで、文書検索、キーワード抽出、文書要約、文書分類等に有用な重要単語のみを特定することができ、これらの処理の精度が向上する。

【００１０】

請求項２に記載の発明は、請求項１に記載の単語出現度計算装置において、前記出現度計算手段は、前記特定の部位を前記文書における見出しとする。

【００１１】

したがって、文字列から抽出した単語の有用度を決定するのに、その単語が所定の文書群の中で出現する全出現数に比して重要な部位である見出しに出現する度合いを計算することで、文書検索、キーワード抽出、文書要約、文書分類等に有用な重要単語のみを特定することができ、これらの処理の精度が向上する。

【００１２】

請求項３に記載の発明は、請求項１に記載の単語出現度計算装置において、前

記出現度計算手段は、前記特定の部位を前記文書における要約とする。

【００１３】

したがって、文字列から抽出した単語の有用度を決定するのに、その単語が所定の文書群の中で出現する全出現数に比して重要な部位である要約に出現する度合いを計算することで、文書検索、キーワード抽出、文書要約、文書分類等に有用な重要単語のみを特定することができ、これらの処理の精度が向上する。

【００１４】

請求項４に記載の発明は、請求項１に記載の単語出現度計算装置において、前記出現度計算手段は、前記特定の部位を前記文書における見出し及び要約とする。

【００１５】

したがって、文字列から抽出した単語の有用度を決定するのに、その単語が所定の文書群の中で出現する全出現数に比して重要な部位である見出し及び要約に出現する度合いを計算することで、文書検索、キーワード抽出、文書要約、文書分類等に有用な重要単語のみを特定することができ、これらの処理の精度が向上する。

【００１６】

請求項５に記載の発明は、請求項２に記載の単語出現度計算装置において、前記出現度計算手段は、“前記文書群で前記単語が前記見出しで出現する文書数／前記文書群で前記単語の出現する全文書数”を計算することにより前記出現の度合いを計算する。

【００１７】

したがって、文字列から抽出した単語が所定の文書群の中で出現する全出現数に比して重要な部位である見出しに出現する度合いを簡易、適切に計算することができる。

【００１８】

請求項６に記載の発明は、請求項３に記載の単語出現度計算装置において、前記出現度計算手段は、“前記文書群で前記単語が前記要約で出現する文書数／前記文書群で前記単語の出現する全文書数”を計算することにより前記出現の度

合いを計算する。

【００１９】

したがって、文字列から抽出した単語が所定の文書群の中で出現する全出現数に比して重要な部位である要約に出現する度合いを簡易、適切に計算することができる。

【００２０】

請求項７に記載の発明は、請求項４に記載の単語出現度計算装置において、前記出現度計算手段は、“前記文書群で前記単語が前記見出し又は前記要約で出現する文書数 ／ 前記文書群で前記単語の出現する全文書数”を計算することにより前記出現の度合いを計算する。

【００２１】

したがって、文字列から抽出した単語が所定の文書群の中で出現する全出現数に比して重要な部位である見出し及び要約に出現する度合いを簡易、適切に計算することができる。

【００２２】

請求項８に記載の発明は、請求項４に記載の単語出現度計算装置において、前記出現度計算手段は、“（単語が見出しで出現する文書数 ／ 単語の出現する全文書数） ＋ （単語が要約で出現する文書数 ／ 単語の出現する全文書数）”を計算することにより前記出現の度合いを計算する。

【００２３】

したがって、文字列から抽出した単語が所定の文書群の中で出現する全出現数に比して重要な部位である見出し及び要約に出現する度合いを簡易、適切に計算することができる。

【００２４】

請求項９に記載の発明は、請求項１～８の何れかの一に記載の単語出現度計算装置において、前記特定の部位の種類について選択を受付ける選択手段を備え、前記出現度計算手段は、この選択された特定の部位における前記出現の度合いを計算する。

【００２５】

したがって、見出し、要約など複数種類の特定の部位から所望のものを選択して、その部位における単語の出現度合いを計算できるので、ユーザーの使い勝手を向上させることができる。

【００２６】

請求項１０に記載の発明は、請求項１～９の何れかの一に記載の単語出現度計算装置と、前記出現の度合いを利用して前記文書から検索語を選出する検索語選出手段と、この選出した検索語に適合する文書を前記文書群から選出する文書選出手段と、を備えている文書検索装置である。

【００２７】

したがって、文書検索の処理精度を向上させることができる。

【００２８】

請求項１１に記載の発明は、請求項１～９の何れかの一に記載の単語出現度計算装置と、前記出現の度合いを利用して前記文書からキーワードを抽出するキーワード抽出手段と、を備えているキーワード抽出装置である。

【００２９】

したがって、キーワード抽出の処理精度を向上させることができる。

【００３０】

請求項１２に記載の発明は、請求項１～９の何れかの一に記載の単語出現度計算装置と、前記出現の度合いを利用して前記文書からキーワードを抽出するキーワード抽出手段と、前記出現の度合いを利用して前記文書から文を抽出して要約文とする要約作成手段と、を備えている文書要約装置である。

【００３１】

したがって、文書要約の処理精度を向上させることができる。

【００３２】

請求項１３に記載の発明は、請求項１～９の何れかの一に記載の単語出現度計算装置と、前記出現の度合いを利用して前記文書から分類キーワードを抽出する分類キーワード抽出手段と、この抽出結果を利用して前記文書群の文書を分類する分類手段と、を備えている文書分類装置である。

【００３３】

したがって、文書分類の処理精度を向上させることができる。

【００３４】

請求項１４に記載の発明は、文字列の入力を受付ける入力手段と、この受付けた文字列から単語を抽出する単語抽出手段と、この抽出した各単語について所定の第１の文書群の各文書における出現の度合いを計算する出現度計算手段と、この出現の度合いを利用して前記文書から検索語を選出する検索語選出手段と、この選出した検索語に適合する文書を前記第１の文書群とは別の第２の文書群から選出する文書選出手段と、を備えている文書検索装置である。

【００３５】

したがって、第２の文書群の文書と異なる文体や語彙や内容領域を持つ検索要求が入力された場合に、単語の有用度を決定するのに、入力した文書と同種の文体や語彙や内容領域を持つ文書からなる第１の文書群のそれぞれの単語の出現する度合いを計算すれば、入力文書から抽出した単語の第１の文書群で出現する度合いが第２の文書群で出現する度合いより大きいものは、有用度を下げることが可能となり、入力文書の同種文書に特有の単語を除くことができ、文書検索、キーワード抽出、文書要約、文書分類等の処理の精度が向上する。

【００３６】

請求項１５に記載の発明は、請求項１～９の何れかの一に記載の単語出現度計算装置の各手段が実行する処理をコンピュータに実行させるコンピュータに読取り可能なプログラムである。

【００３７】

したがって、請求項１～９の何れかの一に記載の発明と同様の作用を奏する。

【００３８】

請求項１６に記載の発明は、請求項１０又は１５に記載の文書検索装置の各手段が実行する処理をコンピュータに実行させるコンピュータに読取り可能なプログラムである。

【００３９】

したがって、請求項１０又は１５に記載の発明と同様の作用を奏する。

【００４０】

請求項１７に記載の発明は、請求項１１に記載のキーワード抽出装置の各手段が実行する処理をコンピュータに実行させるコンピュータに読取り可能なプログラムである。

【００４１】

したがって、請求項１１に記載の発明と同様の作用を奏する。

【００４２】

請求項１８に記載の発明は、請求項１２に記載の文書要約装置の各手段が実行する処理をコンピュータに実行させるコンピュータに読取り可能なプログラムである。

【００４３】

したがって、請求項１２に記載の発明と同様の作用を奏する。

【００４４】

請求項１９に記載の発明は、請求項１３に記載の文書分類装置の各手段が実行する処理をコンピュータに実行させるコンピュータに読取り可能なプログラムである。

【００４５】

したがって、請求項１３に記載の発明と同様の作用を奏する。

【００４６】

請求項２０に記載の発明は、請求項１５～１９の何れかの一に記載のプログラムを記憶している記憶媒体である。

【００４７】

したがって、その記憶しているプログラムにより請求項１５～１９の何れかの一に記載の発明と同様の作用を奏する。

【００４８】

【発明の実施の形態】

［発明の実施の形態］

本発明の一実施の形態を発明の実施の形態１として説明する。

【００４９】

図１は、本実施の形態である文書検索装置１の電気的な接続を示すブロック図

である。図１に示すように、文書検索装置１は、ＰＣなどのコンピュータであり、各種演算を行ない文書検索装置１の各部を集中的に制御するＣＰＵ２と、各種のＲＯＭやＲＡＭからなるメモリ３とが、バス４で接続されている。

【００５０】

バス４には、所定のインターフェイスを介して、ハードディスクなどの磁気記憶装置５と、マウスやキーボードなどで構成される入力装置６と、ＬＣＤやＣＲＴなどの表示装置７と、光ディスクなどの記憶媒体８を読取る記憶媒体読取装置９とが接続され、また、インターネットなどのネットワーク１０と通信を行なう所定の通信インターフェイス１１が接続されている。なお、記憶媒体８としては、ＣＤやＤＶＤなどの光ディスク、光磁気ディスク、フレキシブルディスクなどの各種方式のメディアを用いることができる。また、記憶媒体読取装置９は、具体的には記憶媒体８の種類に応じて光ディスクドライブ、光磁気ディスクドライブ、フレキシブルディスクドライブなどが用いられる。

【００５１】

磁気記憶装置５には、この発明のプログラムを実現する情報変換プログラムが記憶されている。この情報変換プログラムは、記憶媒体８から記憶媒体読取装置９により読取るか、あるいは、インターネットなどのネットワーク１０からダウンロードするなどして、磁気記憶装置５にインストールしたものである。このインストールにより文書検索装置１は動作可能な状態となる。この文書検索プログラムは、特定のアプリケーションソフトの一部をなすものであってもよい。また、所定のＯＳ上で動作するものであってもよい。

【００５２】

図２に示すように、この文書検索装置１をサーバコンピュータ１４として実施し、このサーバコンピュータ１４と端末装置１２とをネットワーク１３を介して接続して、端末装置１２からサーバコンピュータ１４を操作できるようにしてもよい。この場合に、端末装置１２は、パーソナルコンピュータ、携帯情報端末（ＰＤＡ）、携帯電話などの情報処理装置として実施することができる。また、ネットワーク１３は、無線、有線及び放送波のいずれを用いたものでもよく、例えば、ＬＡＮ、ＷＡＮ、インターネット、アナログ電話網、デジタル電話網（ＩＳ

ＤＮ）、ＰＨＳ（パーソナルハンディホンシステム）網、携帯電話網、衛星通信網などを利用することができる。

【００５３】

以下では、文書検索プログラムに基づいて文書検索装置１が行なう処理の内容について説明する。

【００５４】

図３は、文書検索プログラムで実現される文書検索装置１の機能を説明する機能ブロック図である。文書検索装置１は、検索要求となる文章の入力を受付ける検索要求入力部２１、検索語候補を抽出して、その検索語としての有用度を算出する検索語選出部２２、検索語候補の指定部位出現度を計算する指定部位出現度計算部２３、文書選出部２４、文書出力部２５、及び、文書データベース２６等より構成される。文書データベース２６は磁気記憶装置５に構築されるものであっても、文書検索装置１の外部に構築されるものであってもよい。

【００５５】

図４は、文書検索プログラムに基づいて文書検索装置１が実行する処理のフローチャートである。まず、検索要求入力部２１により、ユーザーがキーボード等で検索要求となる文章の文字列を入力する（ステップＳ１）。ステップＳ１により入力手段を実現する。この例では、「Ａ社は、昨日、新しいプリンター AcmePrinter を発売した。」という新聞記事からの引用文を検索要求として入力したものとして説明する。

【００５６】

かかる入力があると（ステップＳ１のＹ）、検索語選出部２２は、入力された文章の文字列を所定の単語辞書により形態素解析して単語に分解する（ステップＳ２）。さらに、用意された不要語表に、この抽出した単語が登録されていれば不要語として削除して、残りの単語を検索語候補とする（ステップＳ３）。例えば、上の検索要求なら、「は」や「を」や「した」が不要語として削除され、「Ａ社」「昨日」「新しい」「プリンター」「AcmePrinter」「発売」が検索語候補として残る。このステップＳ２，Ｓ３により単語抽出手段を実現している。

【００５７】

さらに、検索語選出部２２は、この各検索語候補について、検索語としての有用度を算出する。これには、例えば、以下の（１）式を用いることができる。

【００５８】

検索語の有用度＝単語の重み　……　（１）

ここで、「単語の重み」は、一般的には、

“log（全文書数 ／ 単語の出現文書数）”により求めることができる。すなわち、文書データベース２６に登録されている文書群の中で出現文書数の少ない単語は、有用であるとみなす。

【００５９】

しかし、この文書検索装置１では、指定部位出現度計算部２３が、それぞれの単語が検索対象文書である文書データベース２６の文書群の文書中で出現する部位（文書中の「見出し」に出現するか、「要約」に出現するか、など）に着目し、その単語が指定の重要部位に出現する度合い（指定部位出現度）を単語の有用度に反映させる。

【００６０】

例えば、文書の「見出し」を指定部位とした場合、指定部位出現度計算部２３は、

指定部位出現度

＝ 単語が見出しで出現する文書数 ／ 単語の出現する全文書数

……　（２）

により、指定部位出現度を計算する。

【００６１】

あるいは、文書の「要約」を指定部位とした場合には、

指定部位出現度

＝ 単語が要約で出現する文書数 ／ 単語の出現する全文書数

……　（３）

となる。

【００６２】

あるいは、文書中の「見出し」及び「要約」の両方を指定部位とした場合は、

指定部位出現度

＝ 単語が見出し又は要約で出現する文書数 ／ 単語の出現する全文書数

…… （４）

としてもよい。

【００６３】

さらに、上記（２）式と（３）式とを組み合わせ、

指定部位出現度

＝ （単語が見出しで出現する文書数 ／ 単語の出現する全文書数）

＋ （単語が要約で出現する文書数 ／ 単語の出現する全文書数）

…… （５）

としてもよい。

【００６４】

何れの手段でも、指定部位出現度を計算することにより、文書中における指定の重要部位で多く使われる単語を見分けることができる。その前提として、文書データベース２６の電子化されている各文書について「見出し」「要約」などの各部分の範囲が文書中のどこからどこまでであるかを示すデータを持っているか、あるいは、各文書について「見出し」「要約」などの各部分ごとに各単語の出現数のデータを予め備えている必要がある。

【００６５】

このようにして指定部位出現度計算部２３が検索語候補の指定部位出現度を計算すると（ステップＳ４）、検索語選出部２２は、指定部位出現度計算部２３の算出した検索語候補の指定部位出現度を利用して検索語候補の有用度を計算して、検索語を抽出する（ステップＳ５）。ステップＳ４により出現度計算手段を、ステップＳ５により検索語選出手段を実現している。そして、ステップＳ１～ステップＳ４の機能により単語出現度計算装置を実現している。

【００６６】

すなわち、（１）式から、

検索語の有用度＝単語の重み×指定部位出現度 …… （６）

となる。

【００６７】

あるいは、検索要求文章が長い場合には、

検索語の有用度

＝単語の重み×指定部位出現度×検索要求文章内での出現回数

…… （７）

のように計算することもできる。

【００６８】

このように、指定部位出現度を利用することにより、文書中の指定の重要部位で多く使われる単語を優先させることができる。

【００６９】

この点につき、前述の文例で具体的に説明する。この文例は、「Ａ社は、昨日、新しいプリンター AcmePrinter を発売した。」であり、「Ａ社」「昨日」「新しい」「プリンター」「AcmePrinter」「発売」が検索語候補であった。

【００７０】

下記の表１は、この各検索語候補である「単語」について、文書データベース２６に登録されている文書群中で出現する文書の数を「出現文書数」、その中でも文書の見出しで出現する文書の数を「見出しでの出現文書数」、文書の要約で出現する文書の数を「要約での出現文書数」として例示したものである。

【００７１】

【表１】

単語	見出しでの出現文書数	要約での出現文書数	出現文書数
A社	22	22	30
昨日	0	10	16
新しい	2	8	24
AcmePrinter	8	8	12
発売	20	26	32

【００７２】

この例において、（１）式で単語の有用度を計算すると、「昨日」は有用度が高いとみなされるが、（６）式で指定の重要部位に出現する度合いを利用して単

語の有用度を計算するなら、こういった単語の有用度は低く計算されることがわかる。

【００７３】

このように各検索語候補について有用度がもとまったら、ステップＳ５において、検索語選出部２２は、有用度の高い順に検索語候補を並べ、例えば、その上位１０位を検索語として選出する。

【００７４】

そして、文書選出部２４は、検索語選出部２２が選出した検索語を用いて、文書データベース２６を検索し、適合する文書を選定する（ステップＳ６）。ステップＳ６により文書選出手段を実現している。

【００７５】

この選定された適合文書は、文書出力部２５へ渡される。文書出力部２５は、文書選出部２４で選出した適合文書を、検索結果として出力する（ステップＳ７）。

【００７６】

また、部位種類指定部２７は、指定部位出現度計算部２３が前述のように指定部位出現度を計算する際の文書中の部位の種類（「見出し」か、「要約」か、あるいはその両方か）の選択を、ユーザーから受付ける。そして、この選択に応じて、指定部位出現度計算部２３は（２）～（５）式の何れかにより指定部位出現度を計算する。

【００７７】

［発明の実施の形態２］

別の実施の形態を発明の実施の形態２として説明する。

【００７８】

図５は、この実施の形態である文書検索装置１の機能ブロック図である。この文書検索装置１のハードウエア構成は、図１、図２を参照して説明した発明の実施の形態１の場合と同様であり、詳細な説明は省略する。

【００７９】

この文書検索装置１が実施の形態１と相違するのは、文書群（第１の文書群）

を登録した第１の文書データベース３１と、別の文書群（第２の文書群）を登録した第２の文書データベース３２とを取り扱うこと、及び、指定部位出現度計算部２３に代えてデータベース出現度計算部３３を備えていることである。

【００８０】

第１の文書データベース３１、第２の文書データベース３２は、磁気記憶装置５に構築されていても、文書検索装置１の外部に構築されていてもよい。第２の文書データベース３２は前述の文書データベース２６に相当するもので、検索対象文書からなる文書データベースである。第１の文書データベース３１は、検索要求と同種の文体や語彙や内容領域を持つ文書からなる文書データベースである。この例では、第２の文書データベース３２には特許公報の文書群がおさめられ、第１の文書データベース３１には新聞記事の文書群がおさめられているものとする。

【００８１】

図６は、文書検索プログラムに基づいて文書検索装置１が実行する処理のフローチャートである。

【００８２】

ステップＳ１１～Ｓ１３の処理は、前述のステップＳ１～Ｓ３と同様である。ステップＳ１１により入力手段を、ステップＳ１２，Ｓ１３により単語抽出手段を実現している。この例でも、検索要求入力部２１により、「Ａ社は、昨日、新しいプリンター AcmePrinter を発売した。」といった新聞記事からの引用文を入力したものとして説明する。ここでも、「Ａ社」「昨日」「新しい」「プリンター」「AcmePrinter」「発売」が検索語候補として残る。そして、前述と同様に、各検索語候補について検索語としての有用度を（１）式により算出すると、第２の文書データベース３２での出現文書数の少ない単語は、有用であるとみなされることとなる。

【００８３】

しかし、本文書検索装置１では、データベース出現度計算部３３が、それぞれの単語が、検索要求文書と同種の文体や語彙や内容領域を持つ文書からなる第１の文書データベース３１で出現する頻度にも着目し、その頻度と、同じ単語が第

２の文書データベース３２で出現する頻度との違いの度合い（データベース出現度）を、有用度に反映させる。そのために、まず、データベース出現度を計算する（ステップＳ１４）。ステップＳ１４により出現度計算手段を実現している。また、ステップＳ１１～Ｓ１４の機能により単語出現頻度計算装置を実現している。

【００８４】

例えば、データベース出現度計算部３３は、データベース出現度の算出のために、

データベース出現度

＝ 第２の文書データベースでの出現文書数 ／ 第２の文書データベース全文書数

－ 第１の文書データベースでの出現文書数 ／ 第１の文書データベース全文書数

（ただし、値が負になる場合は、データベース出現度を０とする）

…… （８）

のような計算をする。

【００８５】

あるいは、

データベース出現度

＝ （第２の文書データベースでの出現文書数 ／ 第２の文書データベース全文書数） ／ （第１の文書データベースでの出現文書数 ／ 第１の文書データベース全文書数）

（ただし、値が１未満になる場合は、データベース出現度を１とする）

…… （９）

のように計算してもよい。

【００８６】

このようにして、第１の文書データベース３１での単語の出現頻度と、第２の文書データベース３２での単語の出現頻度とを用いてデータベース出現度を計算することにより、第２の文書データベース３２では、比較的使われないが、第１

の文書データベース３１ではよく使われる単語を選ばれにくくすることができる。

【００８７】

そして、検索語選出部２２は、データベース出現度計算部３３の算出するデータベース出現度を利用して単語の有用度を計算し、検索語を抽出する（ステップＳ１５）。

【００８８】

すなわち、（１）式から、

検索語の有用度

＝単語の重み×データベース出現度　　……　（１０）

となる。

【００８９】

この点につき、前述の文例で具体的に説明する。この文例は、「Ａ社は、昨日、新しいプリンター AcmePrinter を発売した。」であり、「Ａ社」「昨日」「新しい」「プリンター」「AcmePrinter」「発売」が検索語候補であった。

【００９０】

下記の表２は、この各検索語候補である「単語」について、第１の文書データベース３１に登録されている文書群中で出現する文書の数を「第１の文書データベースでの出現文書数」、第２の文書データベース３２に登録されている文書群中で出現する文書の数を「第２の文書データベースでの出現文書数」として例示したものである。

【００９１】

【表２】

単語	第１のデータベースでの出現文書数	第２のデータベースでの出現文書数
リコー	30	3
昨日	16	0
新しい	24	18
プリンター	12	10
AcmePrinter	6	0
発売	32	5

【００９２】

この例において、例えば（１）式で単語の有用度を計算すると、「Ａ社」や「発売」といった単語は有用度が高いとみなされるが、（１０）式で単語の有用度を計算するなら、こういった単語の有用度は低く計算されることがわかる。

【００９３】

ステップＳ１５では、このように各検索語候補について有用度がもとまったら、検索語選出部２２が、有用度の高い順に検索語候補をならべ、例えば、上位１０位までを検索語として選出する。ステップＳ１５により文書選出手段を実現している。

【００９４】

ステップＳ１６，Ｓ１７の処理については、前述のステップＳ６，Ｓ７と同様であり、ここでは説明を省略する。

【００９５】

なお、この例では、検索要求と検索対象とで文書の種類が異なる場合を例として説明した。すなわち、第１、第２の文書データベース３１、３２に登録されている文書群として新聞と特許公報とを例として挙げて説明した。この他に、同じ種類の文書であっても、検索要求と検索対象とで異なる分野に属する場合（例えば、特許公報であってもＩＰＣ分類が異なる場合など）や、検索要求と検索対象とが異なる著者の文書による場合などにも、この文書検索装置１は有益である。

【００９６】

なお、実施の形態１と実施の形態２とを組み合わせて用いることもできる。す

なわち、単語の出現度をもとめるのに、指定部位出現度計算部２３とデータベース出現度計算部３３を併用するものである。

【００９７】

［発明の実施の形態３］

別の実施の形態を発明の実施の形態３として説明する。

【００９８】

図７は、この実施の形態であるキーワード抽出装置４１の機能ブロック図である。このキーワード抽出装置４１のハードウエア構成は、図１、図２を参照して説明した発明の実施の形態１の場合と同様であり、詳細な説明は省略する。

【００９９】

このキーワード抽出装置４１では、図１のハードウエア構成で、記憶媒体８やネットワーク１０からのダウンロードからインストールしたキーワード抽出プログラムが動作する。そして、キーワード抽出プログラムに基づく処理により、実施の形態１と同様な文書データベース２６を扱い、実施の形態１と同様な機能を有する指定部位出現度計算部２３と、キーワード抽出部４２と、部位種類指定部２７とを実現している。

【０１００】

図８は、キーワード検索プログラムに基づいてキーワード抽出装置４１が実行する処理のフローチャートである。まず、キーワード抽出部４２に、文書が入力されると（ステップＳ２１のＹ）、その文書を対象に前述のステップＳ２，Ｓ３と同様の処理を行なう（ステップＳ２２，Ｓ２３）。これにより、入力文書からキーワード候補となる単語が抽出される。ステップＳ２１により入力手段を、ステップＳ２，Ｓ３により単語抽出手段を実現している。

【０１０１】

指定部位出現度計算部２３は、各キーワード候補の指定部位出現度を、実施の形態１の場合と同様にして計算する（ステップＳ２４）。ステップＳ２４により出現度計算手段を実現している。また、ステップＳ１～Ｓ４により単語出現度計算装置を実施している。

【０１０２】

そして、キーワード抽出部４２は、指定部位出現度計算部２３で算出された指定部位出現度を用いて単語の有用度を実施の形態１の場合と同様に求め、有用度の高い順にキーワード候補を並べて、例えば、上位１０位までをキーワードとして選出する（ステップＳ２５）。ステップＳ２５によりキーワード抽出手段を実現している。

【０１０３】

このようにして、各文書の特徴をあらわすキーワードを的確に抽出することができる。

【０１０４】

［発明の実施の形態４］

別の実施の形態を発明の実施の形態４として説明する。

【０１０５】

図９は、この実施の形態である文書要約装置５１の機能ブロック図である。この文書要約装置５１のハードウエア構成は、図１、図２を参照して説明した発明の実施の形態１の場合と同様であり、詳細な説明は省略する。

【０１０６】

このでは、図１のハードウエア構成で、記憶媒体８やネットワーク１０からのダウンロードからインストールした文書要約プログラムが動作する。そして、文書要約プログラムに基づく処理により、実施の形態３と同様な文書データベース２６を扱い、実施の形態３と同様な機能を有する指定部位出現度計算部２３と、キーワード抽出部４２とを実現している。実施の形態３と相違するのは、後述のような機能を備えた要約作成部５２も実現している点である。

【０１０７】

図１０は、文書要約プログラムに基づいて文書要約装置５１が実行する処理のフローチャートである。ステップＳ３１～Ｓ３４は、前述のステップＳ２１～Ｓ２４と同様の処理である。ステップＳ３１により入力手段を、ステップＳ３２，Ｓ３３により単語抽出手段を、ステップＳ３４により出現度計算手段を、それぞれ実現している。また、ステップＳ３１～Ｓ３４の機能により単語出現度計算装置を実施している。そして、実施の形態３の場合と同様に、キーワード抽出部４

２でキーワードを抽出する（ステップＳ３５）。ステップＳ３５によりキーワード抽出手段を実現している。

【０１０８】

このようにして、各文書の特徴をあらわすキーワードが得られるので、要約作成部５２は、ステップＳ３１で入力された文書から、このキーワードを所定程度多く含んでいる文だけを抽出し（ステップＳ３６）、これらの文からなる文書を要約文として出力する（ステップＳ３７）。例えば、キーワードを多く含む順に上位１０位までの文を抽出することなどが考えられる。ステップＳ３６により要約作成手段を実現している。

【０１０９】

このようにして、要約文を的確に作成することができる。

【０１１０】

［発明の実施の形態５］

別の実施の形態を発明の実施の形態５として説明する。

【０１１１】

図１１は、この実施の形態である文書分類装置６１の機能ブロック図である。この文書分類装置６１のハードウエア構成は、図１、図２を参照して説明した発明の実施の形態１の場合と同様であり、詳細な説明は省略する。

【０１１２】

この文書分類装置６１では、図１のハードウエア構成で、記憶媒体８やネットワーク１０からのダウンロードからインストールした文書分類プログラムが動作する。そして、文書分類プログラムに基づく処理により、実施の形態１と同様な文書データベース２６を扱い、実施の形態１と同様な機能を有する指定部位出現度計算部２３、部位種類指定部２７を実現している。さらに、後述のような機能を備えた分類キーワード選出部６２と、分類部６３も実現している。

【０１１３】

図１２は、文書分類プログラムに基づいて文書分類装置６１が実行する処理のフローチャートである。まず、分類キーワード選出部６２に、文書が入力されると（ステップＳ４１のＹ）、この文書を対象として前述のステップＳ２，Ｓ３と

同様の処理を実行する（ステップＳ４２，Ｓ４３）。このようにして抽出された単語を分類キーワード候補とする。ステップＳ４１により入力手段を、ステップＳ４２，Ｓ４３により単語抽出手段を実現している。

【０１１４】

次に、指定部位出現度計算部２３は、各分類キーワード候補の指定部位出現度を計算する（ステップＳ４４）。ステップＳ４４により出現度計算手段を実現している。また、ステップＳ４１～Ｓ４４の機能により単語出現度計算装置を実施している。

【０１１５】

そして、分類キーワード選出部６２は、算出された指定部位出現度を用いて単語の有用度を実施の形態１の場合と同様に求め、有用度の高い順に分類キーワード候補を並べて、例えば、上位１０位までを分類キーワードとして抽出する（ステップＳ４５）。ステップＳ４５により分類キーワード抽出手段を実現している。

【０１１６】

このようにして文書ごとに選出された分類キーワードに基づいて、分類部６３は、文書を分類する（ステップＳ４６）。ステップＳ４６により分類手段を実現している。これには、例えば、分類キーワードの単語ごとの有用度を要素とするベクトルを作成し、互いの内積を算出して、ベクトル間の距離を求め、距離の近いものどうしを同じ分類とすること等で実現する。これらについては周知の技術であるため、詳細な説明は省略する。このようにして分類された文書が得られる。

【０１１７】

【発明の効果】

請求項１に記載の発明は、文字列から抽出した単語の有用度を決定するのに、それぞれの単語が所定の文書群の中で出現する部位に着目し、その単語が出現する全出現数に比して所定の重要な部位に出現する度合いを計算することで、文書検索、キーワード抽出、文書要約、文書分類等に有用な重要単語のみを特定することができ、これらの処理の精度が向上する。

【０１１８】

請求項２に記載の発明は、請求項１に記載の発明において、文字列から抽出した単語の有用度を決定するのに、その単語が所定の文書群の中で出現する全出現数に比して重要な部位である見出しに出現する度合いを計算することで、文書検索、キーワード抽出、文書要約、文書分類等に有用な重要単語のみを特定することができ、これらの処理の精度が向上する。

【０１１９】

請求項３に記載の発明は、請求項１に記載の発明において、文字列から抽出した単語の有用度を決定するのに、その単語が所定の文書群の中で出現する全出現数に比して重要な部位である要約に出現する度合いを計算することで、文書検索、キーワード抽出、文書要約、文書分類等に有用な重要単語のみを特定することができ、これらの処理の精度が向上する。

【０１２０】

請求項４に記載の発明は、請求項１に記載の発明において、文字列から抽出した単語の有用度を決定するのに、その単語が所定の文書群の中で出現する全出現数に比して重要な部位である見出し及び要約に出現する度合いを計算することで、文書検索、キーワード抽出、文書要約、文書分類等に有用な重要単語のみを特定することができ、これらの処理の精度が向上する。

【０１２１】

請求項５に記載の発明は、請求項２に記載の発明において、文字列から抽出した単語が所定の文書群の中で出現する全出現数に比して重要な部位である見出しに出現する度合いを簡易、適切に計算することができる。

【０１２２】

請求項６に記載の発明は、請求項３に記載の発明において、文字列から抽出した単語が所定の文書群の中で出現する全出現数に比して重要な部位である要約に出現する度合いを簡易、適切に計算することができる。

【０１２３】

請求項７に記載の発明は、請求項４に記載の発明において、文字列から抽出した単語が所定の文書群の中で出現する全出現数に比して重要な部位である見出し

及び要約に出現する度合いを簡易、適切に計算することができる。

【０１２４】

請求項８に記載の発明は、請求項４に記載の発明において、文字列から抽出した単語が所定の文書群の中で出現する全出現数に比して重要な部位である見出し及び要約に出現する度合いを簡易、適切に計算することができる。

【０１２５】

請求項９に記載の発明は、請求項１～８の何れかの一に記載の発明において、見出し、要約など複数種類の特定の部位から所望のものを選択して、その部位における単語の出現度合いを計算できるので、ユーザーの使い勝手を向上させることができる。

【０１２６】

請求項１０に記載の発明は、文書検索の処理精度を向上させることができる。

【０１２７】

請求項１１に記載の発明は、キーワード抽出の処理精度を向上させることができる。

【０１２８】

請求項１２に記載の発明は、文書要約の処理精度を向上させることができる。

【０１２９】

請求項１３に記載の発明は、文書分類の処理精度を向上させることができる。

【０１３０】

請求項１４に記載の発明は、第２の文書群の文書と異なる文体や語彙や内容領域を持つ検索要求が入力された場合に、単語の有用度を決定するのに、入力した文書と同種の文体や語彙や内容領域を持つ文書からなる第１の文書群のそれぞれの単語の出現する度合いを計算すれば、入力文書から抽出した単語の第１の文書群で出現する度合いが第２の文書群で出現する度合いより大きいものは、有用度を下げることが可能となり、入力文書の同種文書に特有の単語を除くことができ、文書検索、キーワード抽出、文書要約、文書分類等の処理の精度が向上する。

【０１３１】

請求項１５～１９の何れかの一に記載の発明は、請求項１～１４の何れかの一

に記載の発明と同様の効果を奏する。

【０１３２】

請求項２０に記載の発明は、請求項１５～１９の何れかの一に記載の発明と同様の効果を奏する。

【図面の簡単な説明】

【図１】

本発明の実施の形態１である文書検索装置の電気的な接続を示すブロック図である。

【図２】

文書検索装置をサーバコンピュータとして端末装置と接続して使用する構成例のブロック図である。

【図３】

文書検索装置の機能ブロック図である。

【図４】

文書検索装置が行なう処理を説明するフローチャートである。

【図５】

本発明の実施の形態２である文書検索装置の機能ブロック図である。

【図６】

文書検索装置が行なう処理を説明するフローチャートである。

【図７】

本発明の実施の形態３であるキーワード抽出装置の機能ブロック図である。

【図８】

キーワード抽出装置が行なう処理を説明するフローチャートである。

【図９】

本発明の実施の形態４である文書要約装置の機能ブロック図である。

【図１０】

文書要約装置が行なう処理を説明するフローチャートである。

【図１１】

本発明の実施の形態５である文書分類装置の機能ブロック図である。

【図１２】

文書分類装置が行なう処理を説明するフローチャートである。

【符号の説明】

１　　文書検索装置

８　　プログラム

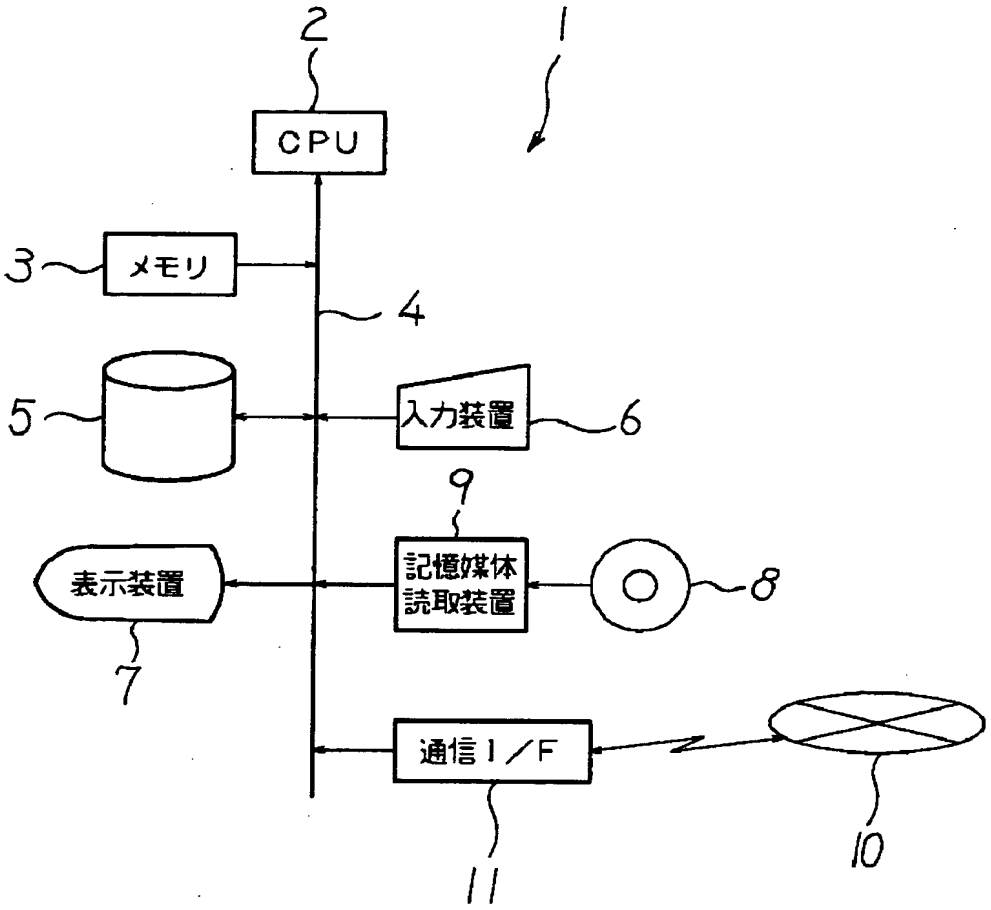
４１　　キーワード抽出装置

５１　　文書要約装置

６１　　文書分類装置

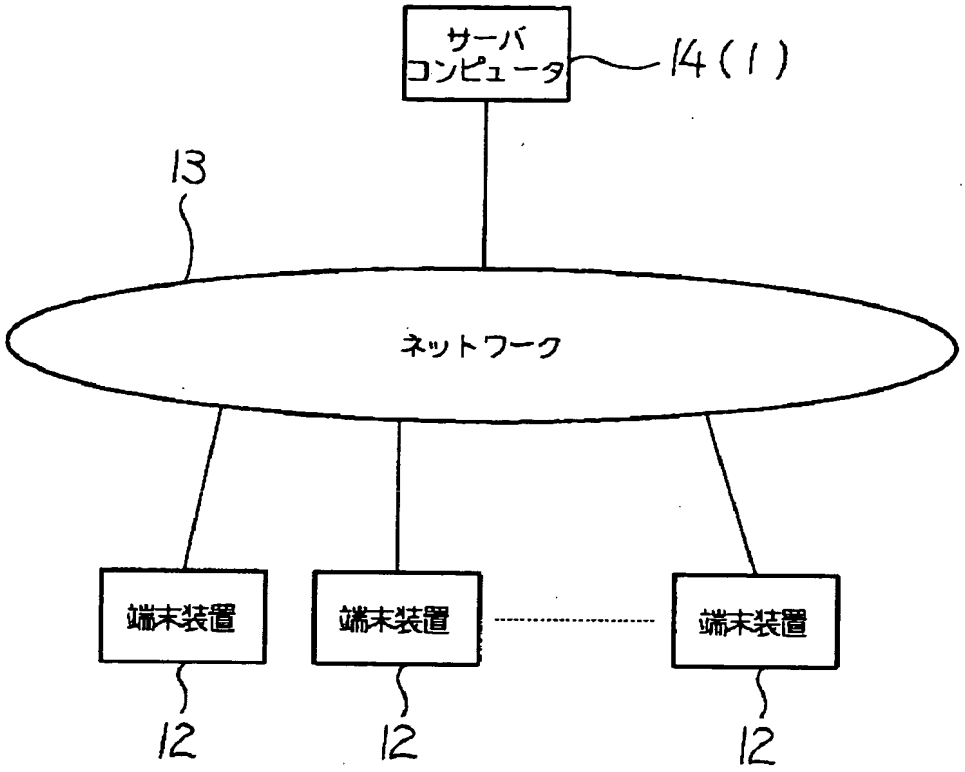
【書類名】　　図面

【図１】

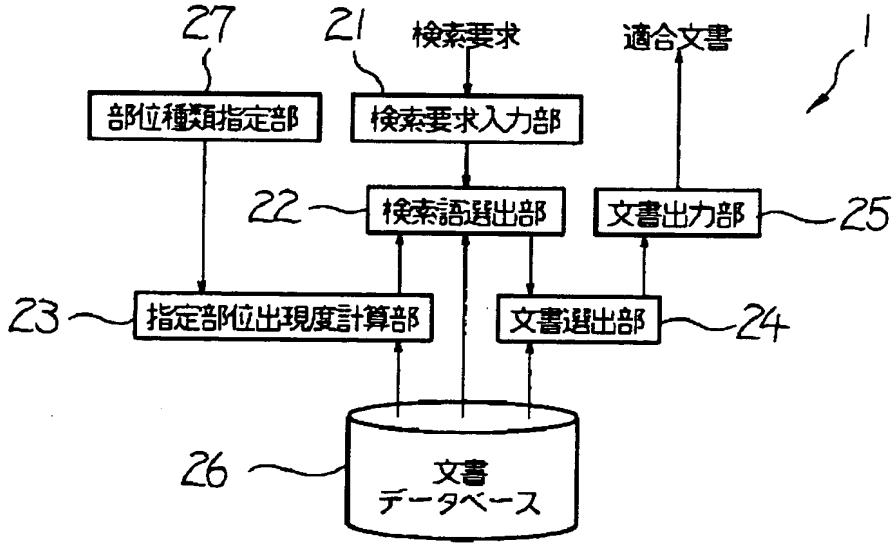


出証特２００３－３０５２８４８

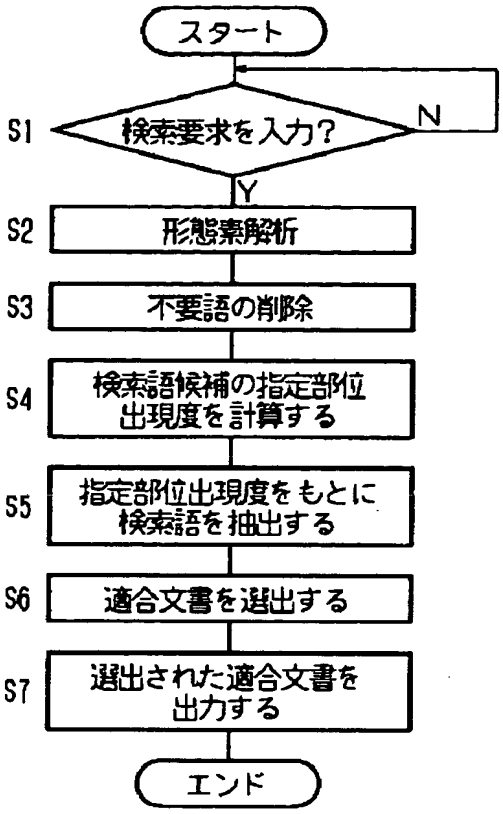
【図２】


サーバ
コンピュータ
14(1)
13
ネットワーク
端末装置
端末装置
端末装置
12
12
12

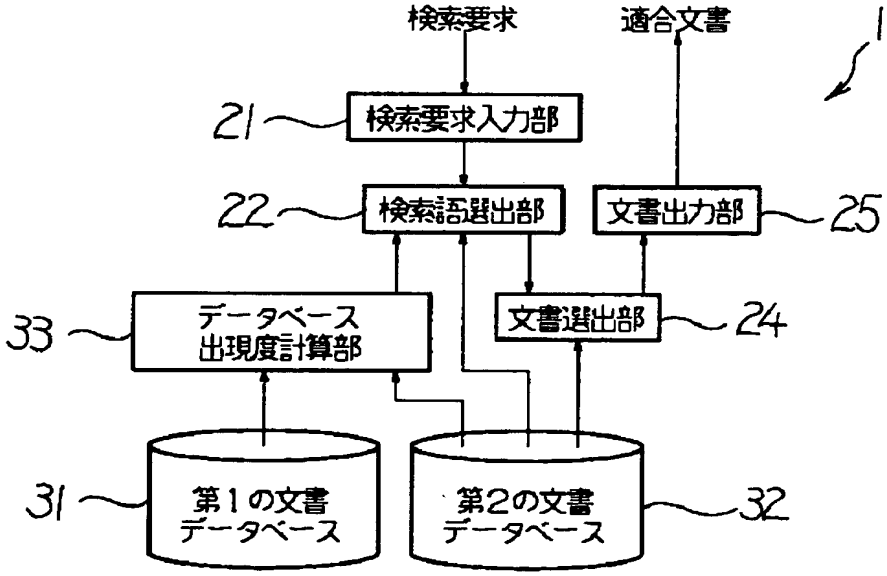
【図３】


27
21
検索要求
適合文書
1
部位種類指定部
検索要求入力部
22
検索語選出部
文書出力部
25
23
指定部位出現度計算部
文書選出部
24
26
文書
データベース

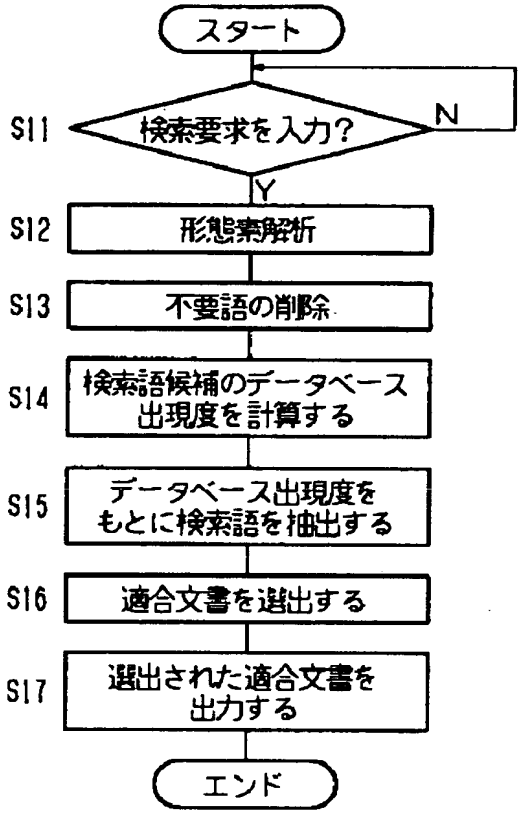
【図４】


スタート
S1
検索要求を入力?
N
Y
S2
形態素解析
S3
不要語の削除
S4
検索語候補の指定部位
出現度を計算する
S5
指定部位出現度をもとに
検索語を抽出する
S6
適合文書を選出する
S7
選出された適合文書を
出力する
エンド

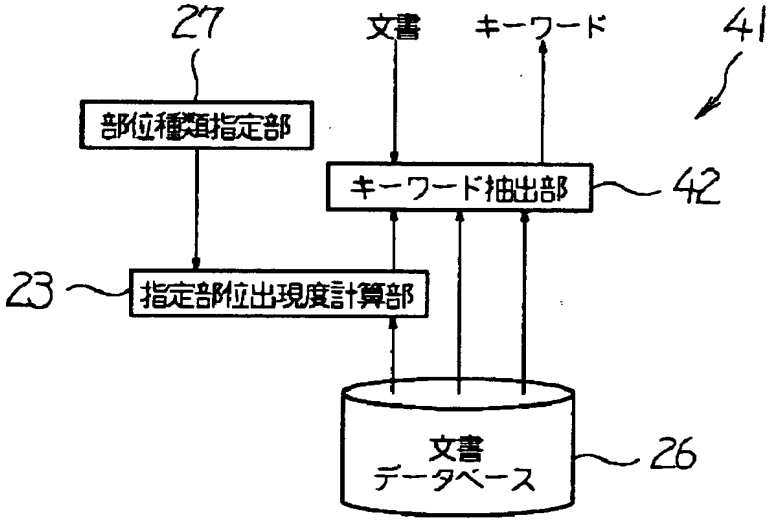
【図５】


検索要求
適合文書
1
21
検索要求入力部
22
検索語選出部
文書出力部
25
33
データベース
出現度計算部
文書選出部
24
31
第1の文書
データベース
第2の文書
データベース
32

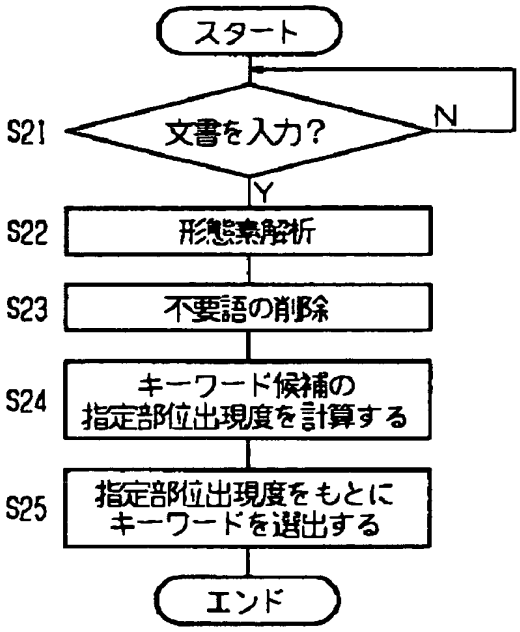
【図６】


スタート
S11
検索要求を入力?
N
Y
S12
形態素解析
S13
不要語の削除
S14
検索語候補のデータベース出現度を計算する
S15
データベース出現度をもとに検索語を抽出する
S16
適合文書を選出する
S17
選出された適合文書を出力する
エンド

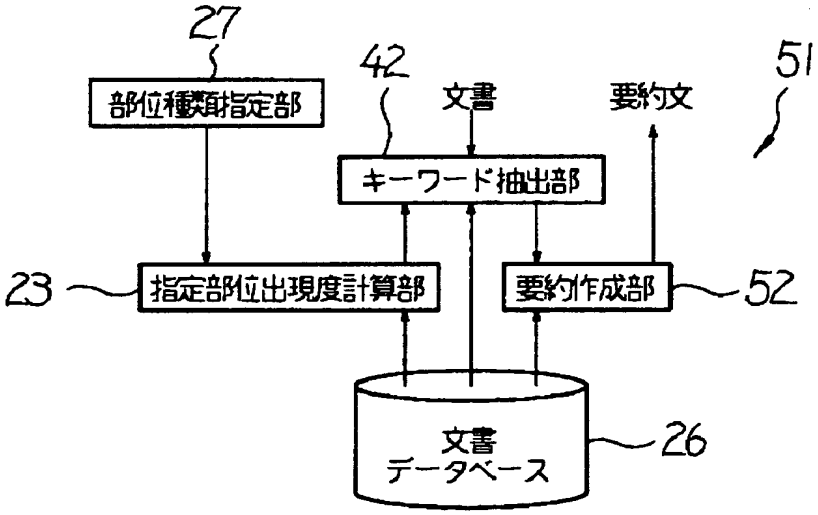
【図７】


27
文書
キーワード
41
部位種類指定部
キーワード抽出部
42
23
指定部位出現度計算部
文書データベース
26

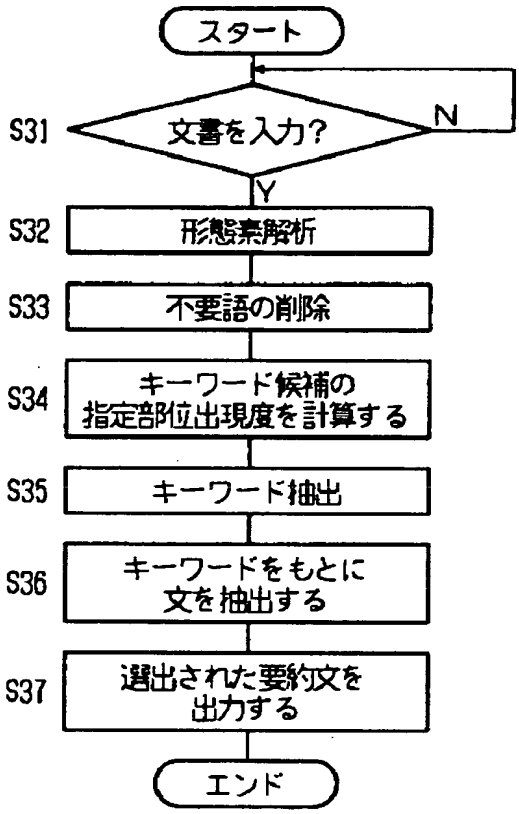
【図８】


スタート
S21
文書を入力？
N
Y
S22
形態素解析
S23
不要語の削除
S24
キーワード候補の
指定部位出現度を計算する
S25
指定部位出現度をもとに
キーワードを選出する
エンド

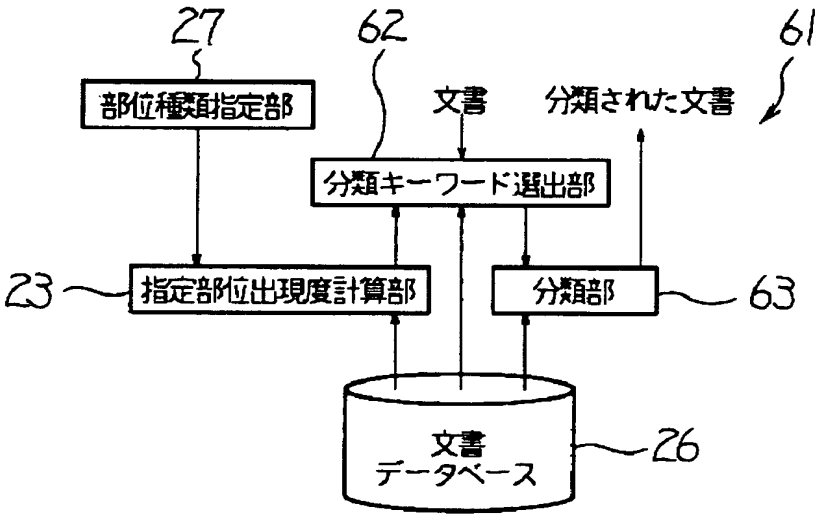
【図９】


27
部位種類指定部
42
文書
要約文
51
キーワード抽出部
23
指定部位出現度計算部
要約作成部
52
文書
データベース
26

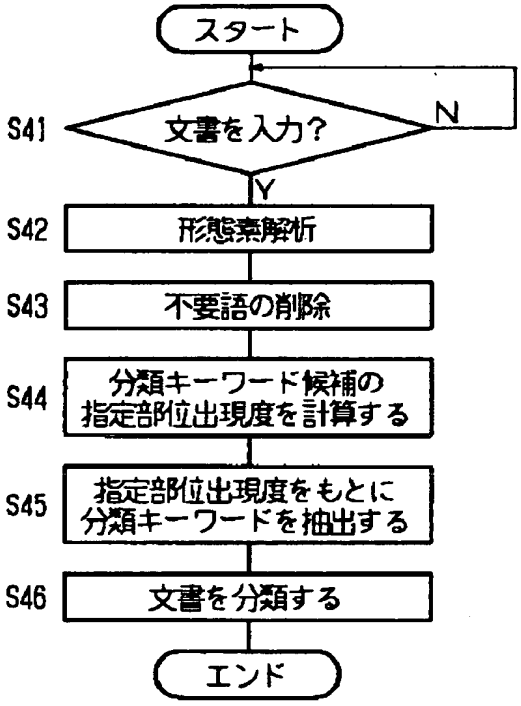
【図１０】


スタート
S31
文書を入力?
N
Y
S32
形態素解析
S33
不要語の削除
S34
キーワード候補の
指定部位出現度を計算する
S35
キーワード抽出
S36
キーワードをもとに
文を抽出する
S37
選出された要約文を
出力する
エンド

【図１１】


27
62
61
部位種類指定部
文書
分類された文書
分類キーワード選出部
23
指定部位出現度計算部
分類部
63
文書
データベース
26

【図１２】


スタート
S41
文書を入力?
N
Y
S42
形態素解析
S43
不要語の削除
S44
分類キーワード候補の
指定部位出現度を計算する
S45
指定部位出現度をもとに
分類キーワードを抽出する
S46
文書を分類する
エンド

【書類名】　　要約書

【要約】

【課題】　長い文書が入力された場合でも、文書検索に有用な重要語のみを選出できるようにして、文書検索の精度を向上する。

【解決手段】　検索要求の文書から不要語を除去した（ステップＳ２，Ｓ３）後の各単語が、検索対象文書である文書中の見出し、要約などの重要部位に出現する度合い（指定部位出現度）を計算する（ステップＳ４）。そして、指定部位出現度をもとに検索語を抽出し（ステップＳ５）、この検索語から適合文書を選出する（ステップＳ６）。

【選択図】　　図４

出 願 人 履 歴 情 報

識別番号　　〔０００００６７４７〕

１．変更年月日　　２００２年　５月１７日
［変更理由］　住所変更
住　所　東京都大田区中馬込１丁目３番６号
氏　名　株式会社リコー

出証特２００３－３０５２８４８